2008 10th Intl. Conf. on Control, Automation, Robotics and Vision
Hanoi, Vietnam, 17–20 December 2008

Robust Fusion Using Boosting and Transduction for Component-Based Face Recognition

Fayin Li and Harry Wechsler
Department of Computer Science
George Mason Univerisity
Fairfax, VA 22030, USA
fayin.li@gmail.com, wechsler@gmu.edu

Massimo Tistarelli
Computer Vision Laboratory
DAP, Universita di Sassari
Alghero 07041, Italy
tista@uniss.it

***Abstract*—Face recognition performance depends upon the input variability as encountered during biometric data capture including occlusion and disguise. The challenge met in this paper is to expand the scope and utility of biometrics by discarding unwarranted assumptions regarding the completeness and quality of the data captured. Towards that end we propose a model-free and non-parametric component-based face recognition strategy with robust decisions for data fusion that are driven by transduction and boosting. The conceptual framework draws support throughout from discriminative methods using likelihood ratios. It links at the conceptual level forensics and biometrics, while at the implementation level it links the Bayesian framework and statistical learning theory (SLT). Feature selection of local patch instances and their corresponding high-order combinations, exemplar-based clustering (of patches) as components including the sharing (of exemplars) among components, and finally decision-making regarding authentication using boosting driven by components that play the role of weak-learners, are implemented in a similar fashion using transduction driven by a strangeness measure akin to typicality. The feasibility, reliability, and utility of the proposed open set face recognition architecture vis-$\grave{a}$-vis adverse image capture conditions are illustrated using FRGC data. The potential for future developments concludes the paper.**

***Index Terms*—biometrics, boosting, component-based recognition, data fusion, face recognition, disguise, forensics, k-nearest neighbor, likelihood ratio, margin, Neyman-Pearson, occlusion, open set recognition, surveillance, transduction, strangeness, typicality.**

I. Introduction

The biometrics processing space can be thought of as an n-D space with the axes describing variability along dimensions that relate to the data acquisition conditions encountered during enrollment and testing. The axes describe among others geometry for imaging, temporal change, and un-cooperative subjects together with impostors vis-$\grave{a}$-vis occlusion and disguise ("denial and deception"). The challenge met here is to expand the scope and utility of biometrics by discarding unwarranted assumptions (located at the origin of the n-D space) regarding the completeness and quality of the biometric data captured. Image variability and correspondence using precise alignment are major challenges for object recognition, in general, and face recognition, in particular. Component-based face recognition facilitates authentication because it does not seek for explicit invariance. Instead, it handles variability using component-based configurations that are flexible enough to compensate for limited pose changes, if any, and limited occlusion and disguises. The next but obvious question is how to define and derive the components ("parts"). Similar to neural Darwinism the components are emergent local representations that are the result of competitive processes that seek to make legitimate associations between appearance and their (non-accidental) coincidences, on one side, and class labels, on the other side. Feed-forward (cortical) architectures provide the wetware that supports such processes in an incremental fashion. The feed-forward aspect is a limited version of the latency and evidence accumulation concepts [1] reiterated by psychophysical experiments (see result 18) [2]. Evidence accumulation involves a steady progression in the way that visual information is processed and analyzed. "This comes from bandwidth requirements and the need for an early and fast impression, categorization or recognition of the input. Much of the processing required to achieve such a phenomenal amount of computation in such a short time must be based on essentially feed-forward mechanisms." Progressive processing squares well with sparse coding driven by suspicious (non-accidental) coincidences [3] and has been shown to "generalize well to novel views of the same face [for identification]" [4]. The outline for the paper follows. Sect. II provides background and motivation for the scope of the biometric effort. Sects III and IV address complementary issues related to forensics and discriminative methods. Sects V and VI describe the building blocks using transduction for learning and boosting for ensemble methods. Sects VII and VIII are about representation and learning and prediction, respectively. Sect. IX is about experiments, Sect. X discusses the findings and their implication, while Sect. XI concludes the paper.

II. Background

The working hypothesis for the (large) face recognition evaluations carried out so far has not been particularly concerned with the very possibility that subjects seek to deny and/or foil their true biometric signatures. The subjects wanted and/or under surveillance, however, are well motivated to hinder the acquisition of their biometrics. Recent large scale face recognition evaluations, e.g., FRVT2002, FRGC, FRVT2006, still do not consider occlusion (avoiding detection) and disguise (masquerading) for testing purposes. Our own evaluation

studies have shown that the performance displayed by well know face recognition benchmark methods, e.g., PCA and PCA + LDA ("Fisherfaces"), deteriorates significantly as a result of disguise [5]. Occlusion and disguise are not always deliberate. Examples for accidental occlusion are characteristic of crowded environments, e.g., CCTV, when only face components / poses of faces are visible from time to time and not necessarily in the right sequence. Subjects ("targets") can appear and disappear as time progresses and the presence of any face is not necessarily continuous across (video) frames. Some of the CCTV frames could actually be void of any face, while other frames could include occluded or disguised faces from different subjects. The goal is to identify the (CCTV) frames where the same subject ("target"), either known ("enrolled") or unknown, shows up. Enrolled ("familiar") subjects need to be identified as well. This corresponds to the problems of open set face recognition [6] including face selection. Open set face recognition is different from closed set recognition where the assumption is that all the subjects seen were previously enrolled and each authentication requires a forced choice decision.

III. FORENSICS

Gonzales-Rodriguez et al. [7] provide strong motivation from forensic sciences for the evidential and discriminative use of likelihood ratio (LR). They make the case for rigorous quantification of the process leading from evidence (and expert testimony) to decisions. Classical forensic reporting provides only "identification" or "exclusion / elimination" decisions. It has two main drawbacks. The first one is related to the use of subjective thresholds. If the forensic scientist is the one choosing the thresholds, he will be ignoring the prior probabilities related to the case, disregarding the evidence under analysis and usurping the role of the Court in taking the decision, "*... the use of thresholds is in essence a qualification of the acceptable level of reasonable doubt adopted by the expert*" [8]. The second drawback is the large amount of non-reporting or inconclusive cases that this identification / exclusion process has induced. The Bayesian approach's use of the likelihood ratio avoids the above drawbacks. The roles of the forensic scientist and the judge/jury are now clearly separated. What the Court wants to know are the posterior odds in favor of the prosecution proposition (*P*) against the defense (*D*) [posterior odds = LR$\times$ prior odds]. The prior odds concern the Court (background information relative to the case), while the likelihood ratio, which indicates the strength of support from the evidence, is provided by the forensic scientist. The forensic scientist cannot infer the identity of the probe from the analysis of the scientific evidence, but gives the Court the likelihood ratio for the two competing hypothesis (*P* and *D*). The likelihood ratio serves as an indicator of the discriminating power (similar to Tippett plots) for the forensic system, e.g., the face recognition engine, and it can be used to comparatively assess authentication performance. The use of the likelihood ratio has been motivated recently also by specific linkages between biometrics and forensics [9] with the evidence evaluated using a probabilistic framework. Forensic inferences correspond now to authentication, exclusion, or inconclusive outcomes and are based on the strength of biometric (filtering) evidence accrued by prosecution and defense competing against each other. The evidence consists of concordances and discordances for the components making up the facial landscape. The likelihood ratio LR is a quotient of a similarity factor, which supports the evidence that the query sample belongs to a given suspect (assuming that the null hypothesis is made by the prosecution *P*), and a typicality factor, e.g., UBM (Universal Background Model) which quantifies support for the alternative hypothesis made by the defense *D* that the query sample belongs to someone else (see Sect. V for the similarity between LR and the strangeness measure).

IV. DISCRIMINATIVE METHODS

Discriminative methods support practical intelligence. Progressive processing, evidence accumulation, and fast decisions are the hallmarks. There is no time for expensive density estimation, marginalization, and synthesis characteristic of generative methods. There are additional philosophical and linguistic arguments that support the discriminative approach. Philosophically, it has to do with practical reasoning and epistemology, when recalling from Hume, that "all kinds of reasoning consist in nothing but a comparison and a discovery of those relations, either constant or inconstant, which two or more objects bear to each other." The likelihood ratio LR provides straightforward means for discriminative methods using optimal hypothesis testing. Assume that the null "H0" and alternative "H1" hypotheses correspond to impostor $\boldsymbol{i}$ and genuine $\boldsymbol{g}$ subjects, respectively. The probability to reject the null hypothesis, known as the false accept rate (FAR) or type I error, describes the situation when impostors are authenticated as genuine subjects by mistake. The probability for correctly rejecting the null hypothesis (in favor of the alternative hypothesis) is known as the hit or genuine acceptance ("hit") rate (GAR). It defines the power of the test $1-\beta$ with β the type II error when the test fails to reject the null hypothesis when it is false. The Neyman-Pearson (NP) statistical test $\Psi(x)$ tests in an optimal fashion the null hypothesis against the alternative hypothesis, e.g., $P(\Psi(x) = 1|H0) = \alpha, \Psi(x) = 1$ when $f_g(x)/f_i(x) > \tau$, and $\Psi(x) = 0$ when $f_{\mathbf{g}}(x)/f_{\mathbf{i}}(x) < \tau$ with τ some constant. The Neyman-Pearson lemma further says that for some fixed FAR = α one can select the threshold τ such that the $\Psi(x)$ test maximizes GAR and is the most powerful test for testing the null hypothesis against the alternative hypothesis at significance level α. Specific implementations for $\Psi(x)$ during cascade classification are possible and they are driven by strangeness (transduction) (see Sect. V) and boosting (see Sect. VI).

V. TRANSDUCTION

Transduction is a type of local inference ("estimation") that moves from particular(s) to particular(s). In contrast to inductive inference, where one uses empirical data to approximate a functional dependency (the inductive step [that moves from

particular to general] and then uses the dependency learned to evaluate the values of the function at points of interest (the deductive step [that moves from general to particular]), one now directly estimates (using transduction) the values of the function only at the points of interest from the training data [10]. The simplest mathematical realization for transductive inference is the method of *k*-nearest neighbors. The Cover-Hart theorem [11] proves that asymptotically the one nearest neighbor algorithm is bounded above by twice the Bayes' minimum probability of error. This makes the connection between the Bayesian approach and likelihood ratios, on one side, and strangeness (see below) and transduction, on the other side. Transduction seeks to find, from all possible authentications for unknown faces, the one that is most probable according to the gallery of known faces. Face recognition requires (for discrimination purposes) to compare and rank face images according to the way they are different from each other and to rank them accordingly. Scoring and ranking is done using the *strangeness* and *p-values*, which are introduced and explained below. Transduction and Kolmogorov complexity are closely related. Let #(z) be the length of the binary string z and K(z) be its Kolmogorov complexity, which is the length of the smallest program (up to an additive constant) that a Universal Turing Machine needs as input in order to output z. The randomness deficiency D(z) for string z is $D(z) = \#(z) - K(z)$ with D(z) a measure of how random the binary string z is [12]. The larger the randomness deficiency is the more regular and more probable the string z is. Kolmogorov complexity and randomness using MDL (minimum description length) are closely related. Transduction chooses from all the possible labeling for test data the one that yields the largest randomness deficiency, i.e., the most probable labeling. The strangeness measures the lack of typicality for a face component with respect to its true or putative (assumed) identity label and the labels for all the other faces. Formally, the strangeness measure α_i is the (likelihood) ratio of the sum of the *k* nearest neighbor (*k*-nn) distances *d* from the same class *y* divided by the sum of the *k* nearest neighbor (*k*-nn) distances from all the other classes ($\neg y$).

$$\alpha_i = \frac{\sum_{j=1}^{k} d_{ij}^{y}}{\sum_{j=1}^{k} d_{ij}^{\neg y}} \tag{1}$$

The smaller the strangeness, the larger its typicality and the more probable its (putative) label y is. The strangeness facilitates both feature selection (of image patches) (similar to Markov blankets) and variable selection (dimensionality reduction). One finds empirically that the strangeness, classification margin, sample and hypothesis margin, posteriors, and odds are all related via a monotonically non-decreasing function with a small strangeness amounting to a large margin. The likelihood-like definitions for strangeness are intimately related to discriminative methods. The *p*-values available compare the strangeness values to determine the credibility and confidence in the putative classifications made. The *p*-values bear resemblance to their counterparts from statistics but are not the same [13]. They are determined according to the relative rankings of putative authentications against each one of the identity classes known to the gallery using the strangeness. The standard *p*-value construction shown below, where *l* is the cardinality of the training set *T*, constitutes a valid randomness (deficiency) test approximation [14] for some transductive (putative label y) hypothesis

$$p_y(e) = \frac{\#\{i : \alpha_i \geq \alpha_{new}^{y}\}}{l+1} \tag{2}$$

The interpretation for *p*-values is similar to statistical testing of likelihood ratios used to assess the extent to which the biometric data supports or discredits the null hypothesis (for some specific authentication) (see Sect. III). When the null hypothesis is rejected for each identity class known, one declares that the test image lacks mates in the gallery and the identity query is answered with "none of the above." This corresponds to forensic exclusion with the rejection characteristic of open set (face) recognition [6].

VI. BOOSTING

The basic assumption behind boosting is that "weak" learners can be combined to learn any target concept with probability $1-\eta$. Weak learners, usually built around simple features, learn to classify at better than chance (with probability $1/2+\eta$ for $\eta > 0$). AdaBoost [15] works by adaptively and iteratively re-sampling the data to focus learning on samples that the previous weak (learner) classifier could not master, with the relative weights of misclassified samples increased after each iteration. AdaBoost thus involves choosing *T* effective features h_t to serve as weak (learners) classifiers and using them to construct the separating hyper-planes. The mixture of experts or final boosted (stump) strong classifier *H* is

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x) > \frac{1}{2} \sum_{t=1}^{T} \alpha_t \tag{3}$$

with α the reliability or strength of the weak learner. The constant 1/2 comes in because the boundary is located mid-point between 0 and 1. If the negative and positive examples are at -1 and +1 the constant used is 1 rather than 1/2. The goal for AdaBoost is margin optimization with the margin viewed as a measure of confidence or predictive ability. The weights taken by the data samples are related to their margin and explain the AdaBoost's generalization ability. AdaBoost minimizes (using greedy optimization) some risk functional whose minimum defines logistic regression. AdaBoost converges to the posterior distribution of y conditioned on $\boldsymbol{x}$, and the strong but greedy classifier *H* in the limit becomes the log-likelihood ratio test. The same margin can be also induced using the strangeness and this is the approach taken here (see Sect. VII). The multi-class extensions for AdaBoost are AdaBoost.M1 and .M2 the latter used here to learn strong classifiers with the focus now on both difficult samples to recognize and labels hard to discriminate. The use of features or components as weak learners is justified by their apparent simplicity. The drawback for AdaBoost.M1 comes from its expectation that the performance for the weak learners selected is better than chance. When the number of

classes is $k > 2$, the condition on error is, however, hard to be met in practice. AdaBoost.M2 addresses this problem and allows the weak learner to generate instead a set of plausible labels together with their plausibility (not probability), i.e., $[0, 1]^k$. The AdaBoost.M2 version focuses on the incorrect labels that are hard to discriminate. Towards that end, AdaBoost.M2 introduces a pseudo-loss e_t for hypotheses h_t such that for a given distribution D_t one seeks $h_t : x \times Y \rightarrow [0, 1]$ that is better than chance. "The pseudo-loss is computed with respect to a distribution over the set of all pairs of examples and incorrect labels. By manipulating this distribution, the boosting algorithm can focus the weak learner not only on hard-to-classify examples, but more specifically, on the incorrect labels y that are hardest to discriminate" [15]. The use of Neyman-Pearson is complementary to AdaBoost.M2 training and can meet pre-specified hit and false alarm rates during weak learner selection.

VII. REPRESENTATION

Image patches ("features") at different scales and bandwidth channels are extracted. A Gaussian pyramid is built by blurring the original image and image patches are extracted at each level of the pyramid. The local patches extracted encode 1st or 2nd order statistics. The motivation for the 2nd order patches comes from the importance of suspicious coincidences [3], which states that "two candidate feature A and B should be encoded together if the join appearance probability P(A, B) is much greater than P(A)P(B)." The 2nd order patches are extracted from two local regions that neighbor each other. Next one computes a descriptor for each local patch that is highly distinctive yet is invariant to image variability, e.g., illumination and deformations such as facial expressions. The SIFT descriptor [16], which satisfies such requirements, is used to represent the local patches. SIFT provides robustness against both localization errors and geometric distortions. It is further normalized to unit length in order to reduce the sensitivity to image contrast and brightness changes during the testing stage. Feature ("patch instance") selection takes place next.

Fig. 1. Exemplar-Based Face Components

Since background features are distributed uniformly they are relatively strange and are iteratively discarded using iterative backward elimination that approximates Markov blanket filtering [17]. During face detection, i.e., face (foreground) vs. background, one finds for each patch the closest patches from other images that carry the same class label. If there is only one class of objects, patches from additional background only images are used to compute the strangeness. Competition to prototype the face components is unsupervised and employs k-means clustering. Boosting subsequently employs the components to build corresponding strong classifiers for prediction purposes (see Sect. VIII). The components are exemplar-based combinations rather than singletons (see Fig. 1). This leads to both flexibility and redundancy. Flexibility to match what is most conspicuous and redundancy to allow substitutions when patches and/or components are missing or their appearance has changed. Additional motivation comes from the way objects in inferotemporal (IT) cortex are represented using a variety of combinations of active and inactive cortical column ("patches") for individual features [18]. The exemplar-based representation used provides also effective means to share features ("patches") among components [19] and for transfer learning.

VIII. LEARNING AND PREDICTION

The strangeness is the thread to implement both representation and boosting (learning and prediction on classification). The strangeness, which implements the interface between the face representation and boosting, combines the merits of filter and wrapper classification methods. The coefficients and thresholds for the weak learners, including the thresholds needed for open set recognition and rejection are learned using validation images, which are described in terms of components similar to those found during enrollment. The best feature correspondence for each component is sought between a validation and a training face image over the patches defining that component. The strangeness of the best patch found during training is computed for each validation image under all its putative class labels c ($c = 1, \cdots, C$). Assuming M validation images from each class, one derives M positive strangeness values for each class c, and $M(C-1)$ negative strangeness values. The positive and negative strangeness values correspond to the case when the putative label of the validation and training image are the same or not, respectively. The strangeness values are ranked for all the components available, and the best weak learner h_i is the one that maximizes the recognition rate over the whole set of validation images V for some component i and threshold θ_i. Boosting execution is equivalent of cascade classification [20]. A component is chosen as a weak learner on each iteration (see Fig. 2). The level of

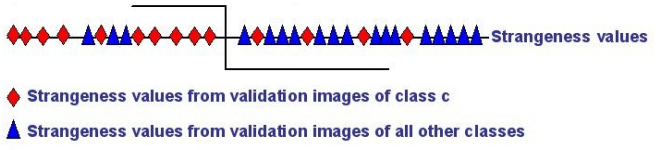

Fig. 2. Learning Weak Learners as Stump Functions

significance determines the scope for the null hypothesi. Different but specific alternatives can be used to minimize Type

II error or equivalently to maximize the power $(1-\beta)$ of the weak learner [21]. During cascade learning each weak learner ("classifier") is trained to achieve (minimum acceptable) hit rate $h = (1-\beta)$ and (maximum acceptable) false alarm rate α. Upon completion, boosting yields the strong classifier $H(x)$, which is a collection of discriminative components playing the role of weak learners. The hit rate after T iterations (see Sect. VI) is h_T and the false alarm α_T.

IX. EXPERIMENTS

The results obtained confirmed first several known psychophysical results [5], among them Result 5 "that of the different facial features, eyebrows were indeed found most important for face detection" (categorization layer 1: face vs. background with Caltech 101 database) using transduction and boosting. The face images in Caltech 101 database have faces as well as clutter background. Faces are not segmented or aligned. The best facial landmark is that component that earned the largest coefficient during boosting. The eye brows are highly discriminative due to their emotive contents, stability and location above a convexity. This makes them less susceptible to shadow and illumination changes. Biometric experiments using the interplay between transduction and boosting were then performed on frontal faces collected at the University of Notre Dame (UND) during 2002-2003, and now part of the FRGC face image database [22]. The experiments are functionally similar to those using multiple samples for face recognition. The face images were acquired under varying illumination (I) (uncontrolled lighting conditions) and/or with varying facial expressions (E). There is also temporal (T) variation as the face images were acquired during different sessions over a one year period. We sampled 200 subjects from the data base; for each one of them there are 48 (frontal) images of which 16 were acquired in an uncontrolled ($I\&E\&T$) environment. The local patches are extracted and the corresponding SIFT descriptors are computed at five scales using $N_s = 5$. Each face is represented by $P = 43N_s = 215$ components described using five feature (patches) exemplars. Using symmetry the number of components comes down to $P = 26N_s = 130$. For each subject, we randomly select 12 images as training set, another 12 images as the validation set and the remaining 24 images as testing set. Euclidian distance is used to compute the strangeness. The top-1 rank identification rates using 1st order patches and strangeness based boosting were 97.5% and 97.9 without and with symmetry, respectively. The corresponding rates using both 1st and 2nd order patches were 98.1% and 98.9%, respectively. Test images were then modified to simulate occlusion. A circle region with radius r is randomly chosen across the face image, the content of which is either set to zero or filled with random pixel values in $[0, 255]$. On the average the recognition rate decreases when the radius of occluded region increases but it does not drop too much. The occluded regions are randomly chosen and the performance observed is very stable when the occluded regions are not too large. The next experiment considered the case when the occluded regions are fixed, e.g., eyes, nose, and mouth, and symmetry is used. The performance is almost the same when one eye is occluded and the other one is available. The occlusion of nose affects the performance more than the mouth and eyes. This is consistent with the relative distribution found for the face components' coefficients and with our earlier findings regarding the importance of the nose for asymmetric faces [23]. Note the nose relevance for second categorization layer ("identification") vs. eye brow importance (discussed earlier) for first categorization layer ("detection").

X. DISCUSSION

One can expand on the thesis put forward by Barlow (1989) regarding suspicious coincidences and their impact on image representations and association codes. Towards that end Balas and Sinha [24] have argued that "rather than relying exclusively on traditional edge-based image representations, it may be useful to also employ region-based strategies that can compare noncontiguous image regions." They further show that "under certain circumstances, comparisons [using dissociated dipole operators] between spatially disjoint image regions are, on average, more valuable for recognition than features that measure local contrast." This leads to the obvious observation that one can and should learn "optimal" sets of regions comparisons for recognizing faces across varying pose and illumination. The choices made on such combinations (during the feature selection stage) amount to "rewiring" operators that connect among lower level operators, usually local ones. This corresponds to a higher processing and competitive stage for the feed-forward and layered architecture. As a result the repertoire of feature now ranges over local, global, and non-local (disjoint) operators ("filters"). Ordinal rather than absolute codes are also possible to gain invariance to small changes in inter-region contrast [24]. The components are clusters described as exemplar-based collections of representative (local or disjoint rewired) patches. Disjoint and "rewired" patches contain more diagnostic information and are expected to perform best for expression, self-occlusion, and varying angle and pose variability. Small-scale local features emerge and are found suitable for recognition under varying illumination. This is in agreement with the optimality of gradient-based features for such tasks [24]. The multi-feature and rewired based representations and exemplar-based components provide added flexibility and should lead to enhanced authentication performance. The mode-free and non-parametric approach presented throughout this paper has handled so far only frontal images possibly affected by adverse data capture conditions. One can expand, however, on the feed-forward architecture to include pose as another dimension that emanates from the origin of the n-D biometric processing space and needs to be addressed (see Sect. I). Layered categorization still starts with face detection but now it seeks for one of three possible poses using boosting driven by relevant components. The poses contemplated are left, frontal, and right. Patches and components are now described using an extended vocabulary of "rewired" operators, both quantitative and qualitative in design (see above).

XI. Conclusion

Biometrics cannot continue to assume that the personal signatures used for face authentication are accurate, complete, constant, and time-invariant. Most clients are indeed legitimate and honest. They have nothing to hide, and have all the incentives to cooperate. The purpose of biometrics, however, is to provide security from impostors seeking to breach security and/or from un-cooperative subjects. Impostors are well motivated to interfere with the proper acquisition of their biometric signatures, and will do their best to hide and/or alter the information needed for their authentication. This paper expands the operational scope for biometrics and addresses situations that involve adverse data capture conditions. The approach taken is realized using boosting and transduction that work together to implement feed-forward (hierarchical) competitive architectures that support component-based (face) recognition strategies. The conceptual framework comes from forensic sciences, the Bayesian framework using the likelihood ratio (LR) and cohorts for discriminative methods, and statistical learning theory (SLT) for hypothesis testing and weak learner selection during boosting. The common thread throughout is the strangeness. It helps with both feature and weak learner selection. The feasibility, reliability, and utility of the proposed open set face recognition architecture vis--vis adverse image capture conditions were illustrated using FRGC data. Venues for future research include open set face selection for video sequences to detect and authenticate subjects whose appearance is sporadic across CCTV frames; and to expand the scope for decision-level fusion to asynchronous multi-sensory data integration. Additional challenges and venues for future research include real-world visual search and categorization (VSC) [25] within complex scenes and the small size problem.

References

[1] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," Nature, vol. 381, pp. 520-522, 1996.

[2] P. Sinha et al., "Face recognition by humans: nineteen results all computer vision researchers should know about," Proc. IEEE, vol. 94, no. 11, pp.1948-1962, 2006.

[3] H. B. Barlow, "Unsupervised learning," Neural Computation, vol. 1, pp. 295-311, 1989.

[4] A. Delorme and S. Thorpe, "Face identification using one spike per neuron: resistance to image degradation," Neural Networks, vol. 14, pp. 795-803, 2001.

[5] H. Lai, V. Ramnathan, and H. Wechsler, "Reliable face recognition using adaptive and robust correlation filters," Computer Vision and Image Understanding (to appear), 2008.

[6] F. Li and H. Wechsler, "Open set face recognition using transduction," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 27, no. 11, pp. 1686-1698, 2005.

[7] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition," IEEE Trans. on Audio, Speech and Language Processing, vol. 15, no. 7, pp. 2104-2115, 2007.

[8] C. Champed and D. Meuwly, "The inference of identity in forensic speaker recognition," Speech Communication, vol. 31, pp. 193-203, 2000.

[9] D. Dessimoz and C. Champod, "Linkages between biometrics and forensic science," Handbook of Biometrics, Anil K. Jain et al., Eds., Springer, 2008.

[10] V. Vapnik, Statistical Learning Theory, Springer, 1998.

[11] T. M. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. on Info. Theory, vol. IT-13, pp. 21-27, 1967.

[12] V. Vovk, A. Gammerman, and C. Saunders, "Machine learning application of algorithmic randomness," 16th Int. Conf. on Machine Learning (ICML), Bled, Slovenia, 1999.

[13] S. S. Ho and H. Wechsler, "Query by transduction," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 30, no. 9, pp. 1557-1571, 2008.

[14] T. Melluish, C. Saunders, A. Gammerman, and V. Vovk, "The typicalness framework: A comparison with the Bayesian approach," TR-S, Royal Holloway ollege, Univ. of London, 2001.

[15] Y. Freund and R. E. Shapire, "Experiments with a new boosting algorithm," 13th Int. Conf. on Machine Learning (ICML), pp. 148 - 156, Bari, Italy, 1996.

[16] D. G. Lowe, "Distinctive image features from scale - invariant key points," Int. Journal of Computer Vision, vol. 60, no. 2, pp.91 - 110, 2004.

[17] D. Koller and M. Sahami, "Toward optimal feature selection," 13th Int. Conf. on Machine Learning (ICML), Bari, Italy, 1996.

[18] K. Tsunoda, Y. Yamane, M. Nishizaki, and M. Tanifuji, "Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns," Nature neuroscience, vol. 4, no. 8, pp. 832-838, 2001.

[19] A. Torralba, K. P.Murphy, and W. T.Freeman,"Sharing visual features for multiclass and multiview object detection, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 29, no. 5, pp. 854-869, 2007.

[20] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," CVPR, Kauai, Hawaii, 2001.

[21] R. O.Duda, P. E.Hart, and D. G.Sork, Pattern Classification (2nd ed.), Wiley, 2000.

[22] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," CVPR, New York, NY, 2005.

[23] S. Gutta and H. Wechsler, "Face recognition using asymmetric faces," 1st Int. Conf. on Biometric Authentication, Hong Kong, China, 2004.

[24] B. J. Balas and P. Sinha, "Region-based representations for face recognition," ACM Transactions on Applied Perception, vol. 3, no. 4, pp. 354-375, 2006.

[25] J. D. Smith, J. S. Redford, L. C. Ghent, and D. A. Washburn, "Visual search and the collapse of categorization," J. of Experimental Psychology: General, vol. 134, no. 4, pp. 443-460, 2005.